Canonical Correspondence Analysis as an Approximation to Gaussian Ordination

Kimberly Welsh Johnson[1] and Naomi Altman[2]

[1]Scirex Corporation, Bloomingdale, Illinois 60108, U.S.A.
[2]Biometrics Unit, Cornell University, Ithaca, New York 14853, U.S.A.

ABSTRACT

Canonical Correspondence Analysis is an approximation to maximum likelihood estimation for Gaussian ordination under certain restrictions of the ordination model (Ter Braak, 1987b). Species tolerances must be equal, and species maxima must be equal or at least independent of the location of the optima. These assumptions are often violated in practice.

This paper develops graphical displays to explore how well species abundances approximate Gaussian curves along the derived environmental axes. As well, a simulation study was performed to determine how well Canonical Correspondence Analysis recovered the true axes when the Gaussian model for species abundance is correct, but the assumptions about the tolerances, maxima and location of optima are violated. The methods were applied to an analysis of an observational study conducted on a fen in the Black Hills of South Dakota.

1. Introduction

Gaussian ordination is a method for deriving underlying environmental axes, using information on species abundance and environmental variables, based on the idea that expected abundance for a species will achieve a maximum under ideal environmental conditions and will decline smoothly away from the ideal. Ter Braak (1987b) showed that under restrictions on the Gaussian ordination model, Canonical Correspondence Analysis (CCA) can be used to approximate certain maximum estimators (MLEs) for Gaussian ordination. The objective is to identify the main patterns associating species information and the environment in which the species exists.

Gaussian ordination models species abundance in each environment as a Poisson random variable with mean described by a scaled Gaussian function of the environmental axes. The axes, ideal environment, scaling factor and width of the Gaussian function are computed by maximum

likelihood. Unfortunately, Gaussian ordination is extremely computationally demanding, so finding an approximation less costly to compute is desirable.

Correspondence Analysis (CA) constructs a set of orthogonal latent variables that most fully explains the variation in the species abundances. CCA restricts the set of constructed axes to linear combinations of observed variables. This paper studies the robustness of CCA as an approximation to Gaussian ordination. As well, a graphical tool is developed to enable the researcher to understand the relationship between species abundances and the derived environmental axes. The goal is to assist the researcher in understanding the ecologically reasonable conditions under which CCA will satisfactorily approximate Maximum Likelihood Estimation for Gaussian ordination, and to allow the researcher to determine whether the species abundances appear to follow the Gaussian model.

2. Statistical Model and Theory

2.1 *Notation*

The following lists the notation that is used in CCA.

$\mathbf{Y} = [y_{ki}]$ species by site matrix, of abundances where $k = 1,\ldots,m$ species and $i = 1,\ldots,n$ sites

$\mathbf{y}_{k+} = \{y_{k+}\}$ = column vector with kth entry the abundance for species k totaled across all sites

$\mathbf{y}_{+i} = \{y_{+i}\}$ = row vector with ith entry the abundance for site i totaled across all species

y_{++} = grand abundance total

$\mathbf{M} = diag(y_{k+})$ = diagonal matrix of species abundances

$\mathbf{N} = diag(y_{+i})$ = diagonal matrix of site abundances

$\mathbf{Z} = [z_{ji}]$ environmental variables by site matrix, where i designates site and $j = 1,\ldots,q$ designates environmental variables

$\overline{\mathbf{Z}} = [\overline{\mathbf{z}}_{ji}]$ the matrix of centered environmental variables computed by $\overline{\mathbf{z}}_{ij} = \mathbf{z}_{ji} - \dfrac{\mathbf{z}_{j+}\mathbf{y}_{+i}}{y_{++}}$

$\mathbf{u} = \{u_k\}$ vector of species ideal environment (estimated optima)

$\mathbf{c} = [c_j]$ canonical coefficients (weights) of the environmental variables

$\mathbf{x} = \{x_i\} = \overline{\mathbf{Z}}'\mathbf{c}$, column vector of site scores, linear combination of environmental variables

2.2 *Model*

Ter Braak (1986, 1987b) shows that if the species abundance data follow a Poisson distribution with the following mean (the Gaussian ordination model),

$$E(y_{ki}) = m_k \exp\left[\frac{-0.5(x_i - u_k)^2}{t_k^2}\right]$$

t_k = species' tolerance
u_k = species' optimum environment
m_k = species' maxima abundance

then CCA approximates Gaussian ordination under the following assumptions on the parameters:

1. The values of the linear combination of the measured environmental variables (site scores) are distributed homogeneously across the entire range of the occurrence of the species along the axis.
2. The values of the species optima (species scores) are distributed homogeneously over an interval about the site scores that is large compared to the tolerances.
3. The species tolerances are equal, or if not equal at least independent of the optima.
4. The species maxima are equal, or if not equal at least independent of the optima.

Even if the Gaussian model is not appropriate, CCA can be extremely useful as a summary technique to relate a set of species variables to a set of environmental variables.

The matrix formulation for CCA is: find $\mathbf{x} = \overline{\mathbf{Z}}'\mathbf{c}$, to maximize the dispersion:

$$\delta = \frac{\mathbf{x}'\mathbf{Y}'\mathbf{M}^{-1}\mathbf{Y}\mathbf{x}}{\mathbf{x}'\mathbf{N}\mathbf{x}} = \frac{\mathbf{c}'\overline{\mathbf{Z}}\mathbf{Y}'\mathbf{M}^{-1}\mathbf{Y}\overline{\mathbf{Z}}'\mathbf{c}}{\mathbf{c}'\overline{\mathbf{Z}}\mathbf{N}\overline{\mathbf{Z}}'\mathbf{c}}$$

which measures the variability of species abundance across the sites, with respect to linear combinations of environmental variables.

Solutions of CCA can be derived from the eigenvalue equation:

$$\overline{\mathbf{Z}}\mathbf{Y}'\mathbf{M}^{-1}\mathbf{Y}\overline{\mathbf{Z}}'\mathbf{c} = \lambda\overline{\mathbf{Z}}\mathbf{N}\overline{\mathbf{Z}}'\mathbf{c}$$

or alternatively:

$$(\overline{\mathbf{Z}}\mathbf{N}\overline{\mathbf{Z}}')^{-1}\overline{\mathbf{Z}}\mathbf{Y}'\mathbf{M}^{-1}\mathbf{Y}'\overline{\mathbf{Z}}\mathbf{c} = \lambda\mathbf{c}$$

where $\mathbf{c}$ is an eigenvector and λ is an eigenvalue..

CCA uses an iterative algorithm which selects a unique linear combination of environmental variables called the site scores (or canonical axis) that maximizes the dispersion δ of the estimated species optima. Subsequent axes can be computed orthogonal to previous axes. The

maximum number of ordination axes that can be extracted is equal to the number of measured environmental variables. However, we are generally interested in reducing the dimension to 4 or less in order to give a succinct description of the association between the environmental variables and species variables.

The algorithm also calculates the eigenvalues and canonical coefficients associated with each axis. The iterative process uses a multiple regression of the site scores on the measured environmental variables. During each iteration, the fitted values are computed and these make up the new site scores to carry through the next step. This process is repeated until convergence. These scores are the first ordination axis of CCA and the corresponding eigenvalue is the maximum dispersion, δ.

The multiple correlation of the site scores on the environmental variables is termed the species-environment correlation. It is a measure of the association between the species and the environment (Ter Braak, 1987b).

The coefficients of the linear combination of the environmental variables comprising the ordination axis are called the canonical coefficients. They indicate the relative weight of each variable in the axis.

2.3 *Parameters of Interest*

CCA calculates an approximation to the MLEs of the species optima, the canonical coefficients, and the site scores, which can be used to find other useful estimators (Ter Braak, 1986, 1987a), but not the species maxima or tolerances. Table 1 gives a detailed description of the six groups of useful parameter estimates generated by CCA.

Table 1
Parameters of Interest in CCA
(3 eigen vectors extracted)

Parameter of Interest	Symbol	Description
Canonical Eigenvalues for Ordination Axes	$\lambda_a \quad a = 1,2,3$	First three eigenvalues obtained by CCA
Canonical Coefficients for Ordination Axes	$\mathbf{c_1} = c_{1a}, c_{1a}, c_{1a}$ $\mathbf{c_2} = c_{2a}, c_{2a}, c_{2a}$ $\mathbf{c_3} = c_{3a}, c_{3a}, c_{3a}$	Best weights of each of the environmental variables in each ordination axis.
Species-Environment Correlations for Axes	ρ_a	Multiple Correlation of the site scores on the environmental variables
Percentage Species Variance Explained by Ordination Axes	$\sigma_{s,a} = \frac{\lambda_a}{\theta_\Sigma}$	Measure of the amount of the species variance accounted for by the corresponding ordination axis of CCA
Percentage Species-Environment Covariance Explained by Axes	$\sigma_{e,a} = \frac{\lambda_a}{\lambda_\Sigma}$	Measure of the amount of the species-environmental covariance accounted for by the corresponding ordination axis of CCA
Intra-set Correlations	$\eta_{aj} = \rho_j \nu_{za,xa}$ $j = 1,2,3$	Measure of the rate of change in species s abundance per unit change in the correspond environmental variable x

Note:
θ_Σ is the sum of all the eigenvalues of CA

λ_Σ is the sum of all the eigenvalues of CCA

$\nu_{za,xa}$ is the correlation between the environmental variables and the site score

2.4 *Diagnostic Graphs*

The output of CCA does not give the researcher a means of assessing the distribution of species abundances along the environmental gradients or the fit of the Gaussian ordination model. LOWESS (Cleveland, 1979) curves can be used to smooth the species abundances against the derived axes to gives a readily interpreted graphical display of the relationship. The smoothed

data can first show if the species abundances appear to follow a Gaussian response model and whether the species have similar maxima and similar tolerances.

3. Simulation

Data were generated for ten species abundances and three environmental variables on fifty plots, for a total of 500 possible species abundance values and 150 environmental variable measurements for each of the 11 simulation conditions listed in Table 2.

The simulated data values were analyzed using CA and CCA. Each simulation condition had 100 runs. Boxplots were used to display summary statistics for different simulation conditions.

All data were generated using the computer program GAUSS. Three environmental variables $\mathbf{z_1}, \mathbf{z_2}$, and $\mathbf{z_3}$ were generated as independent U(0,1) for each plot and for each run of each simulation. The true environmental gradient, $\mathbf{x}$, was constructed from a linear combination of the generated environmental variables, using $\mathbf{x} = 60 + 50\mathbf{z_1}$ for all simulations.

For all simulation conditions, the species abundances were generated from a Poisson distribution with mean dependent on the individual species optimum, maximum, tolerance, and value on the true environmental gradient.

The location of each species optimum on the true environmental gradient was kept constant for the entire duration of the simulation experiment. Each species' individual optimum was assigned a value from 50 to 140, so that species 1 had an optimum value of 50, species 2 had an optimum of 60, and so on up to an optimum value of 140 for species 10.

3.1 *Experimental Design*

The distribution of species maxima differed for six of the simulation conditions. There were a total of five different combinations of species maxima: equal maxima, moderately unequal maxima (independent of optima), severely unequal maxima (independent of optima), moderately unequal maxima (dependent on optima), and severely unequal maxima (dependent on optima).

The species tolerances also differed for six of the simulation conditions. As with the species maxima, there were a total of five combinations of species tolerances: equal tolerances, moderately unequal tolerances (independent of optima), severely unequal tolerances (independent of optima), moderately unequal tolerances (dependent on optima), severely unequal tolerances (dependent on optima).

Only the main factor simulations were run along with two "interesting" interactions. This reduced the number of simulation conditions to eleven combinations shown in Table 2.

Table 2
Simulation Conditions

1.	equal species maxima and tolerances (Control)
2.	moderately unequal species maxima, varying by a factor of two (independent of optima), equal species tolerances
3.	severely unequal species maxima, varying by a factor of ten (independent of optima), equal species tolerances
4.	moderately unequal species maxima, varying by a factor of two (dependent on optima), equal species tolerances
5.	severely unequal species maxima, varying by a factor of ten (dependent on optima), equal species tolerances
6.	equal species maxima, moderately unequal species tolerances, varying by a factor of three (independent of optima).
7.	equal species maxima, severely unequal species tolerances, varying by a factor of ten (independent of optima)
8.	equal species maxima, moderately unequal species tolerances, varying by a factor of three (dependent on optima)
9.	equal species maxima, severely unequal species tolerances, varying by a factor of ten (dependent on optima)
10.	severely unequal maxima, severely unequal tolerances, both varying by a factor of ten (both independent of optima)
11.	severely unequal maxima, severely unequal tolerances, both varying by a factor of ten (both dependent on optima)

3.2 *Boxplots of Parameter Estimates*

Figures 1 through 6 are boxplots of selected parameter estimates under the 11 simulation conditions. For all the boxplots of the parameter estimates, the boxplot of the control simulation (simulation 1 - all assumptions met) was highlighted in gray. A dotted line was drawn at the parameter value (were the environmental axes recovered without error) in each graph.

In the simulation experiment, the canonical coefficients came out to be a multiple of 3.5 of the vector (1, 0, 0) for the first axis. Since this vector can be scaled to any multiple, the multiple was not considered important. Subsequently, in Figure 1, a line is drawn at 3.5 to represent the parameter value.

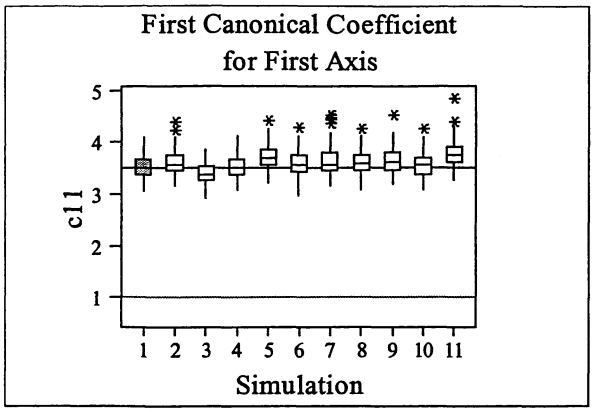

Figure 1: Boxplots for Variable c_{11}

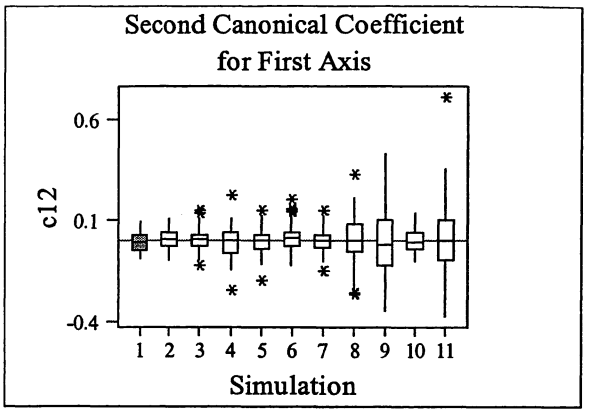

Figure 2: Boxplots for Variable c_{12}

In Figures 1 and 2 it can be seen that for the simulations where the maxima and tolerances were independent of the optima (simulations 1 through 3, and 6 through 8) CCA estimated the parameters well with the median close to the parameter value and small variability. As the tolerances became increasingly unequal and dependent on the optima, the variance of the estimate became larger. When the tolerances were unequal and dependent on the optima, the estimator was centered at the true parameter value but had high variability.

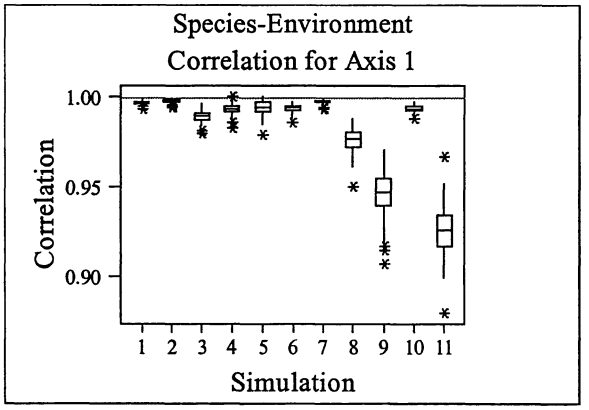

Figure 3: Boxplots for Variable ρ_1

Figure 3 shows boxplots for the parameter estimates for the species-environment correlation. The analysis underestimated this parameter if the tolerances were unequal and dependent upon the optima.

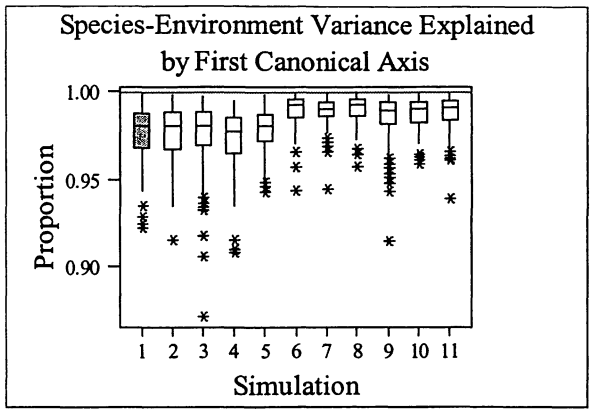

Figure 4: Boxplots for Variable $\sigma_{e,1}$

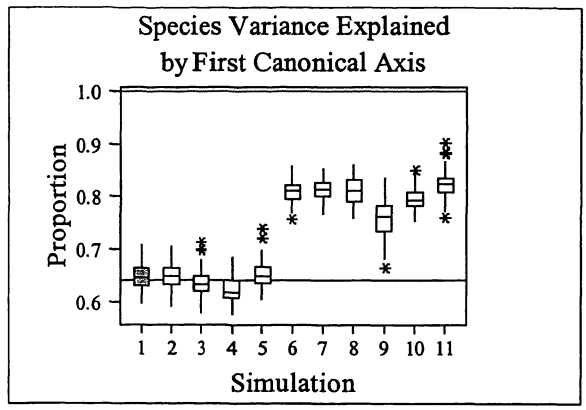

Figure 5: Boxplots for Variable $\sigma_{s,1}$

The species-environment covariance explained by the first axis (Figure 4) was about the same across all simulations. The species variance explained by the first axis (Figure 5) only performed consistently for the first five simulations when the tolerances were held equal. In other situations it appeared to overestimate the parameter.

The estimate of the species variance explained is computed by taking the eigenvalue from the first axis of CCA and dividing it by the sum of all the eigenvalues from CA. In other words it is the ratio of the variance explained by the measured environmental variables in axis 1 to the variance explained by the theoretical environmental axes of CA (the theoretical environmental gradient that most fully explains the variation in the species occurrences). Although the data were generated by a linear combination of the environmental variables, the Poisson variation of the data ensure that the percent variance explained is less than 100%. This makes the parameter estimates of $\sigma_{s,1}$ less than 1. It appears from simulation 1 that almost 65% of the species variation is due to the observed environment, while the other 35% is due to random variation. Overall the estimate performed consistently if the tolerances were equal.

The intra-set correlations performed much like the canonical coefficients. Figure 6 shows that when the tolerances were equal (simulations 1 to 5) the median of the parameter estimates was close to the parameter and the variability was low. Once the tolerances were unequal, the variability increased, especially if the tolerances were dependent upon the optima.

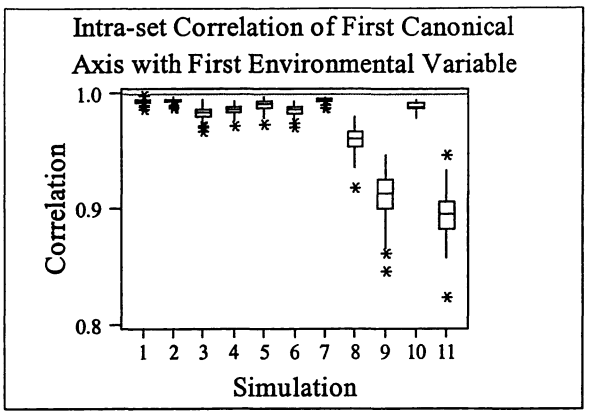

Figure 6: Boxplots for Variable η_{11}

The intra-set correlations performed much like the canonical coefficients. Figure 6 shows that when the tolerances were equal (simulations 1 to 5) the median of the parameter estimates was close to the parameter and the variability was low. Once the tolerances were unequal, the variability increased, especially if the tolerances were dependent upon the optima.

All simulation conditions where the tolerances were held equal (simulations 2 - 5) performed similarly to the control simulation (simulation 1). Once the tolerances were varied (simulations 6 - 11), the parameter estimates tended to estimate incorrectly. The percentage of species-environmental covariance explained by each axis gave the most consistent estimation. The intra-set correlations and canonical coefficients were centered correctly over all the iterations, but the variability was high (especially when no relationship existed). When the tolerance were severely unequal, the correlation between the canonical axes and the environmental axes was underestimated and the percentage of species variance explained was not centered correctly.

3.3 *Graphical Analysis*

For each simulation, the ten generated species abundances were smoothed using the LOWESS smoothing algorithm against the derived environmental gradient. The Poisson intensity curves were also smoothed against the derived environmental gradient for each species. The key for each species in each graph is outlined in Table 3.

Table 3
Key for LOWESS Smooths and Poisson Intensity Curve Graphs

Species	Line Appearance	Line Color
1	Solid	Black
2	Dash	Black
3	Dot	Black
4	1 Dash, 1 Dot	Black
5	1 Dash, 2 Dots	Black
6	1 Dash, 3 Dots	Black
7	Long Dash	Black
8	Solid	Gray
9	Dash	Gray
10	Dot	Gray

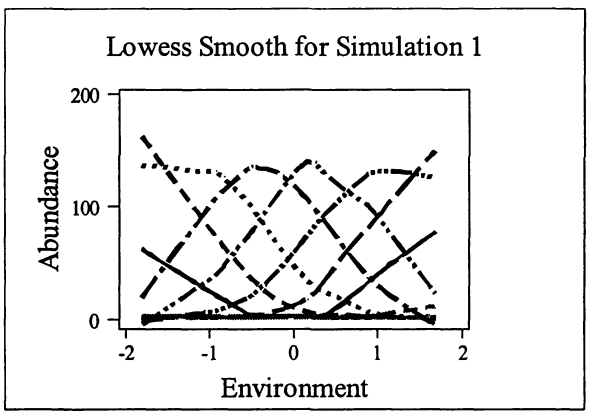

Figure 7: LOWESS Smooth for Observed Abundances (Simulation 1)

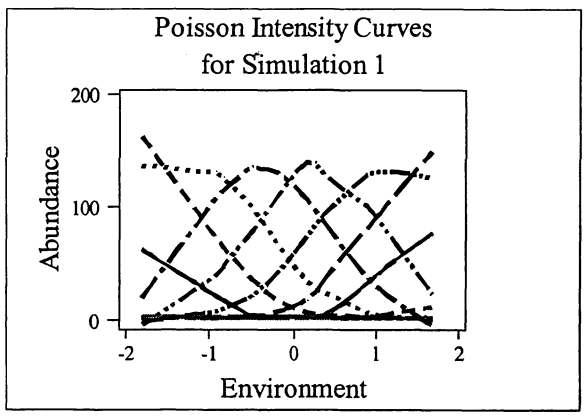

Figure 8: LOWESS Smooth for Expected Abundances (Simulation 1)

Figure 7 displays the LOWESS smooths of observed abundance for one iteration of simulation 1. The smooth for simulation 1 showed that the species had equal maxima and equal tolerances, since each mound was roughly the same height and shape. Because the optima were generated equally spaced, the mounds were also fairly equally spaced apart. Figure 8 displays the smooths of the abundances generated from a Poisson distribution with mean determined by the first equation of section 2.2. Comparing Figures 7 and 8 we can see that the generated species abundances followed the expected Poisson intensity very closely. This type of pattern represents species abundances related to an environmental gradient which satisfied all the conditions under which CCA performed best.

These graphs show that CCA recovered an axis along which the species abundances followed the Gaussian model. Each species abundance and intensity curve had a bell shaped response on the derived environmental gradient. The smooths for rest of the simulations against the CCA axis all look similar to the smooth against the true axis.

5. Macintosh Fen Study

This observational study was conducted during the last three weeks in June, 1994. The area studied was Macintosh Fen, located less than a mile west of Rochford, a small town in the Black Hills of western South Dakota. There are a number of small fens throughout the Black Hills region. The soil in many of these fens has a high iron content. Over the last five to ten years, there has been interest in studying these small fens and in developing information regarding the relationships between the environmental makeup of the soil, especially in regard to the iron content, and the species composition.

Macintosh Fen is located at 44° 7' 11" North, 103° 44' 2" West, at an altitude of 5,460 feet. The fen is long and narrow, and runs in a north/south fashion for about 200 meters along the eastern side of a gravel road. The fen is only 60 meters across at its widest point. The fen itself is flat, and the ground is marshy. A stream runs down the length of the eastern edge, while the road borders the western edge. The ground rises sharply to the east of the stream a few hundred feet in altitude into a mountainous area of Black Hill spruce forest.

5.1 *Experimental Design*

The fen area was divided into four sections from north to south according to the density of brush in the fen. The fen was drier and had more brush in the northern, narrower area. It then widened and opened into wetter ground and sedge cover further south. The sections fell into a rough gradient with section 1 being the furthest north and having the most brush cover to section 4 being the furthest south and having the widest, marshiest, open area.

Thirteen 0.5 by 0.5 meter plots were randomly chosen from each section, making a total of 52 plots sampled. The plots were chosen by mapping the coordinates of the fen map using a grid of 50 by 50 meter squares for a total of 100 possible 0.5 by 0.5 meter plots. Plots were selected at random, and then mapped with a compass and meter tape in the field.

Species and environmental information were measured on each plot. Each different plant species in the plot was classified and the species abundance was recorded. For the environmental information, soil was taken from each plot and sent to a lab to determine the levels of Phosphorus, Potassium, Calcium, Magnesium, Iron, percentage organic matter, and pH.

5.2 *Results*

The final model included all 52 plots. The 7 most abundant species (*Carex rostrata, Poa pratensis, Viola macloskeyi, Fragaria virginiana, Veronica americana, Cerastium nutans, and Sphagnum russowi)* and the 6 of the 7 environmental variables listed above related to soil composition were used. (Magnesium was removed from the model to avoid multicollinearity problems.)

Since section was included for each plot and fell into a gradient from north to south, the analysis was run using section as a covariate. In order to correct for the covariate, CCA first computed an ordination axis for the covariate. It then computed ordination axes, using the environmental data (without the covariate), orthogonal to the covariate axis.

Four axes were extracted in the analysis, but only 2 axes were interpreted, since the additional variance explained axes 3 and 4 were small (Table 4).

Summary

	Axis 1	Axis 2	Axis 3	Axis 4
Eigenvalues	0.197	0.063	0.039	0.013
Species-Environment Correlation	0.683	0.440	0.462	0.227
Cumulative % Variance of: Species Data	13.5	17.8	20.5	21.4
Species-Environment Relation	62.9	83.1	95.6	99.6

Table 4: Summary of CCA results for the first 4 canonical axes.

The four axes only explained about 20% of the total species variance, while they explained almost 100% of the relationship between the species and the environmental variables measured

(Table 4). Most of the variation for both the species data and the species-environmental data was explained by the first two axes (83.1%).

Canonical Coefficients

	pH	P	K	Ca	% OM	Fe
Axis 1	-0.2780	-0.2806	0.4181	-0.1376	-0.3237	-0.9766
Axis 2	0.4627	-0.0667	-0.2274	-1.1540	0.7190	-0.0319

Table 5: Canonical Coefficients of Environmental Variables for the first 2 CCA axes.

The first canonical axis (Table 5) consisted mainly of Iron and Potassium, while the second canonical axis was mainly composed of pH, Organic Matter, and Calcium.

Intra-Set Correlations of Environmental Variables with Axes

	pH	P	K	Ca	% OM	Fe
Axis 1	0.0843	-0.3645	-0.0372	-0.0290	-0.1745	-0.5814
Axis 2	-0.1792	-0.1712	-0.2200	-0.3164	0.1607	-0.0019

Table 6: Correlation between canonical axes and environmental variables.

The intra-set correlations (Table 6) suggest that the first axis was mainly Iron. The second axis had a moderate Calcium intra-set correlation.

5.3 *Conclusions for Canonical Coefficients*

For the Macintosh Fen, the first canonical axis consisted mainly of Iron and Potassium, while the second canonical axis was mainly composed of pH, Organic Matter, and Calcium.

Accounting for variation, the first canonical axis appears to be 0.5Potassium - 1Iron. The environmental gradient which caused the most species variation among the 7 species included in this study (in the first axis) was a combination of Iron and Potassium in a 1:2 negative relationship (2 parts Potassium to -1 part Iron). The second canonical axis was 0.5pH + 0.7Organic Matter - 1Calcium, which made the second axis a roughly 1:1.5:-2 relationship.

In the first environmental axis, Iron seemed to be the most influential environmental variable effecting the species composition, while in the second axis, the effect was made up of mostly Calcium and Organic Matter.

5.4 *Conclusions for Intra-set Correlations*

In the first axis, only Iron had a high correlation with the species composition. This suggests that Potassium did not contribute strongly to the first environmental axis. Considering these correlations, along with the canonical coefficients for the first axis, Iron appeared to be the influential environmental variable for the first axis.

In the second axis, only Calcium had a moderate correlation with the species composition. The scatterplot of Calcium versus pH (Figure 9), shows that pH and Calcium are fairly highly correlated, suggesting that acidity may have played a role in species variation in axis 2. Since pH and Calcium are highly correlated, the intra-set correlation may have been smaller than expected for both pH and Calcium.

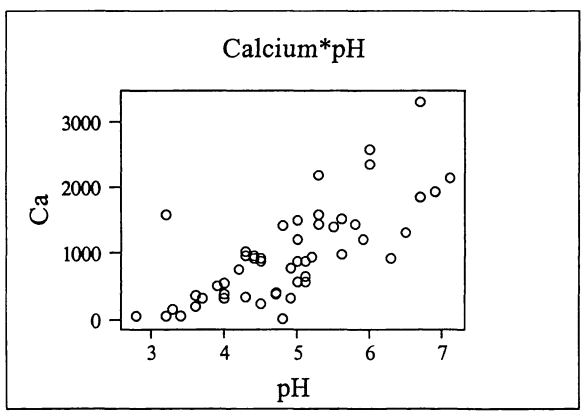

Figure 9: Scatterplot of Calcium versus pH

5.5 *Conclusions for the Species-Environmental Covariance Explained.*

The first axis explained 63% of the variance due to a relationship between the environmental variables measured in the experiment and the species included in the analysis. The second axis explained an additional 20.2%, for a total of 83.1% of the species-environmental covariation explained in the first two axes. Over three quarters of the species-environment relationship from

the variables included in this analysis were due to iron content and calcium content (possibly an indicator of soil acidity).

5.6 *Conclusions for Species Variance Explained*

Even though a large amount of the species-environmental covariation was explained in the first two axes, only a total of 17.8% of the total species variation was explained by the first two axes (13.5% in the first axis and 4.3% in the second). A considerable 80% of the species variation was not explained by any linear combination of the environmental variables measured.

This suggests that while iron and calcium are important factors in species composition in the Macintosh Fen, other important environmental factors that affect species composition have not been measured in this particular study.

5.7 *Diagnostics*

The observed species abundances for the species analyzed in Macintosh Fen were smoothed against the derived environmental gradients. These graphs were then compared with the graphs of the smooths obtained from the simulations to assess how the species from Macintosh Fen are related to one another, and also to determine if the species abundances follow a Gaussian response model. The key for each species in each graph is outlined in Table 7.

Table 7
Key for LOWESS Smooths for Macintosh Fen Species

Species	Line Appearance	Line Color
Carex rostrata	Solid	Black
Poa pratensis	Dash	Black
Viola macloskeyi	Dot	Black
Fragaria virginiana	Long Dash	Black
Veronica americana	Solid	Gray
Cerastium nutans	Dash	Gray
Sphagnum russowii	Dot	Gray

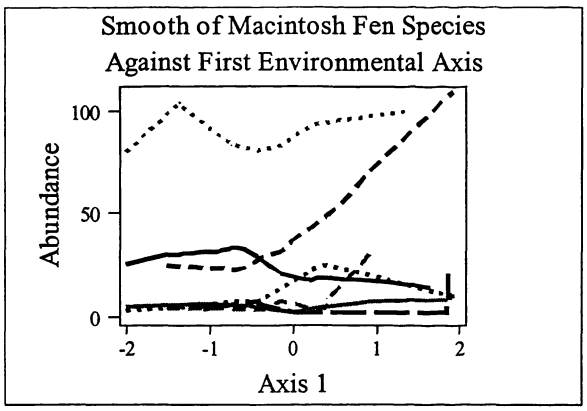

Figure 10: LOWESS Smooth for Fen Abundances
Against First Derived Environmental Axis

It was very difficult to assess from Figure 10 if the species abundances from Macintosh Fen were following a Gaussian response model. *Carex rostrata* and *Viola macloskeyi* had somewhat mound shaped response curves. *Poa pratensis*, *Fragaria virginiana*, and *Cerastium nutans* may be growing at the low end of their tolerance range.

Sphagnum russowii showed an unexpected response curve. From this portion of the graph (Figure 10) it appeared that *Sphagnum russowii* was not unimodal; it almost appeared bi-modal. Or it may have a larger tolerance than any of the other species analyzed. If so, I may have only sampled *Sphagnum russowii* in its most preferred environment, so the graph did not show where the species abundances decline. Without further study, it is impossible to be certain.

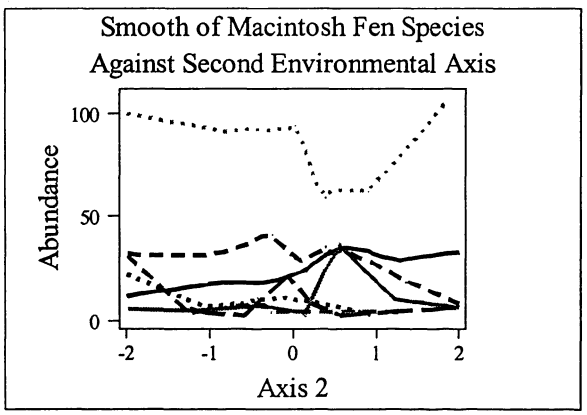

Figure 11: LOWESS Smooth for Fen Abundances Against
Second Derived Environmental Axis

Again, in the smooth against the second environmental axis (Figure 11), *Sphagnum russowii* showed an unexpected response curve, similar to the curve from the first axis. As with the first

axis, it was impossible to draw any definite conclusions without further research into the behavior of *Sphagnum russowii*.

Carex rostrata, *Fragaria virginiana*, and *Veronica americana* in Figure 11 look mound shaped in this axis, with similar spread, implying that these plants may have similar tolerances. *Cerastium nutans* and *Viola macloskeyi* may have occurred in the fen at the high end of their ideal environment in the second axis.

On the whole, the species abundances from Macintosh Fen did not appear to follow a Gaussian response model to these derived environmental gradients.

5.8 *Summary*

After analyzing this data, it appears that the species composition in the Macintosh Fen is moderately affected by both the iron and calcium content of the soil. Species composition is also affected by other environmental variables which were not measured in this study. It appears that soil composition accounts for one tenth to one fifth of the total species variation; whereas other environmental variables may have a large impact on species composition. Some of these factors could be soil moisture, amount of exposure to direct sunlight, minerals in the soil not measured in this study, and other non-mineral environmental factors.

6. Discussion

In general, CCA estimated the parameters well when the tolerances were equal and only the maxima varied. Simulations 8, 9, and 11 seemed to have the worst estimates. These were the simulations in which the tolerances were moderately to severely unequal and dependent upon the optima.

Unequal species tolerances seemed to have a much greater detrimental effect upon the performance of the analysis than unequal species maxima. Even when the maxima were dependent upon the optima, CCA estimated fairly well, as long as the tolerances were held constant. But when the tolerances were varied, even moderately, CCA performed poorly. CCA did not estimate the parameters correctly if the tolerances were dependent upon the optima.

The estimator for the percentage of species-environmental covariance explained by each canonical axis seemed to be the most correctly centered at the parameter value. The canonical

coefficients, intra-set correlations, percentage of species variance explained, and the correlations between the canonical axes and the environmental axes were not as well centered at the parameter value. The canonical coefficients were well centered around the parameter coefficients of the environmental gradient in repeat trials, but the variability was large. The estimator of the canonical coefficients of the environmental gradient was not centered over the parameter value when there was no actual relationship between the axis and the environmental variables. The intra-set correlations performed similarly to the canonical coefficients. When the tolerances were varied, the correlation between the canonical axes and the environmental axes was underestimated in repeat iterations. Worst of all, when the tolerances were varied, the estimator of the percentage of species variance explained was not centered correctly. It gave false information about the amount of species variation due to environment.

CCA performed best when the maxima were within a factor of 10 and the tolerances were equal. The analysis also performed satisfactorily when the tolerances were varied by a moderate amount -- less than a factor of three.

Care should be taken in interpreting both the intra-set correlations and the canonical coefficients since the variability is large. These variables should be loosely interpreted with concentration on the larger picture in the analysis, rather than trying to express the axes in rigid numerical relationships.

The percentage of species variance explained is likely to be overestimated if the tolerances are unequal. In this case, there are also likely to be problems in the estimation of the correlation between the canonical axes and the environmental axes.

LOWESS smooths of the observed species abundances against the derived environmental gradient provide a good diagnostic tool for determining the shape of the relationship between species abundance and the derived axes. If the species abundances are truly Gaussian, CCA will recover an environmental gradient against which the observed abundances are Gaussian, even if the species maxima and tolerances do not follow the assumptions of the model. Looking at the graphs should also give a good general picture of whether the species abundances and tolerances are similar. This is an invaluable tool for determining the fit of a Gaussian ordination model

In conclusion, CCA provides a reasonable approximation to Gaussion ordination when the species being studied adhere to the following basic guidelines: The species have tolerances to

their optimum environment which differ within a factor of 3. The species tolerances are independent of their optimum environment. The species maxima are within a factor of ten from each other.

Under these conditions, the percentage of species variance explained, the percentage of species-environmental covariance explained, and the correlation between the canonical axes and the environmental axes should be well estimated. The intra-set correlations and canonical coefficients should give a good general picture of the relationships being studied.

ACKNOWLEDGMENTS

We would like to thank Dr. Paul Velleman and Dr. Martin Wells for their many suggestions which improved the simulation study, particularly the graphical output. Thanks to Dr. Charles McCulloch for his help with the experimental design of the Macintosh Fen study and for being willing to help with many GAUSS questions. We would also like to thank Dr. Dorothy Chappell for her identification of the plants, and Dr. Cajo ter Braak for his work in the theory of CCA. Much of this work was supported by a teaching assistantship from the Biometrics Unit of Cornell University.

REFERENCES

Braak, C. J. F. ter. (1986). Canonical Correspondence Analysis: A New Eigenvector Technique for Multivariate Direct Gradient Analysis. *Ecology* 67(5): 1167-1179.

Braak, C. J. F. ter. (1987a). The Analysis of Vegetation-Environment Relationships by Canonical Correspondence Analysis. *Vegetatio* 69: 69-77.

Braak, C. J. F. ter. (1987b). Ordination. pp. 91-173 in: *Data Analysis in Community and Landscape Ecology* (R. H. G. Jongman, C. J. F. ter Braak and O. F. R. Van Tongeren, eds.). Wageningen: Pudoc.

Braak, C. J. F. ter. (1987-1992) CANOCO - a FORTRAN program for Canonical Community Ordination. Microcomputer Power, Ithaca, New York, USA.

Cleveland, W.S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *JASA* 74, 829-836.

Chatterjee, Smprit and Ali S. Hadi. (1988). *Sensitivity Analysis in Linear Regression.* New York: John Wiley & Sons, Inc.

GAUSS Ver. 3.0. Aptech Systems, Inc., Maple Valley, WA.

Johnson, Kimberly W. (1996). *Canonical Correspondence Analysis as a Method for Determining Environmental Gradients: A Simulation Experiment Testing Performance.* MS Thesis,Biometrics Unit, Cornell University.

MINITAB Rel. 10. Minitab, Inc., USA.

Ross, Sheldon. (1988). *A First Course in Probability* (third edition). New York: Macmillan Publishing Company.

Searle, Shayle, R. (1982). *Matrix Algebra Useful for Statistics.* New York: John Wiley & Sons, Inc.